

文章编号:1001-9081(2009)01-0269-04

基于信息离散度的 DNA 序列相似性分析

骆嘉伟, 刘芳, 杨华

(湖南大学 计算机与通信学院, 长沙 410082)

(ahfang09@163.com)

摘要: 信息离散度量方法在生物信息处理领域中获得成功地应用, 其基本思想是利用子序列分布差异来表示序列之间的差异, 但是子序列长度的变化对结果的影响较大。文中提出了一种新的基于信息离散度的 DNA 序列相似性分析方法, 利用不同距离的碱基对的联合概率分布差异来表示 DNA 序列之间的差异, 并分析了信息集变化对结果的影响。实验结果表明, 该方法是分析 DNA 序列相似性的简单且有效的工具。当信息集变化时, 相似度较高的序列间的距离值变化很小。

关键词: 碱基对; 完全信息集; 离散度函数; 相似性

中图分类号: TP391 **文献标志码:** A

Analysis of similarity of DNA sequences based on a measure of information discrepancy

LUO Jia-wei, LIU Fang, YANG Hua

(College of Computer and Communication, Hunan University, Changsha Hunan 410082, China)

Abstract: Function of Degree of Disagreement (FDOD) has been successfully applied to the research of bioinformatics. Its performance is greatly impacted by the length of subsequence, when the information included in a sequence is described by its subsequence distribution. In this paper, we proposed a new representation called base-base information set to characterize the sequence, then computed the distance between two sequences by FDOD. The base-base information set comprises the joint probabilities of base pair at a distance of 1 to L, where L is an alterable parameter. Also we analyzed how the distance changes when L is changed. As the experimental result shows, the distance between two sequences is insensitive to the change of L when they are closely similar, so our method is effective for analyzing similarity of sequences.

Key words: base pair; complete information set; Function of Degree of Disagreement (FDOD); similarity

0 引言

DNA 作为绝大多数生物的遗传物质, 其结构和功能的研究一直备受生物学研究者的关注。随着测序水平的提高, DNA 序列的规模和长度都呈指数级增长。确定未知序列的结构和功能除生物学手段外, 计算机辅助手段越来越得到人们的重视。

序列相似性分析的主要目的在于阐明序列之间的同源关系以及从已知序列预测未知序列的结构和功能^[1]。相似性分析的主要方法是序列比对, 取代矩阵和空位罚分直接影响相似性分值, 选取不同的参数会得到不同的计算结果, 所以, 针对不同的研究目标和对象应该构建适宜的取代矩阵^[2]。

近年来, 利用图形表示对 DNA 序列进行相似性分析成为了一大热点^[3-6], 其基本思想是将 DNA 序列表示成一条空间曲线, 通过比较曲线间的差异得到序列间的距离。从空间曲线我们可以直观地看出序列是否相似。如果进一步将曲线转化为矩阵, 选取合适的矩阵不变量, 就可以定量地求出序列之间的距离。

信息理论方法也越来越多地被用于序列相似性分析。这种方法只依赖于序列的原始信息, 并不涉及任何主观因素。但以往的信息理论方法在 DNA 序列的研究中并没有取得理

想的结果^[7], 一些研究结果甚至否定了信息理论在序列比较问题上的可用性^[8]。文献[9]提出一种新的信息离散度量方法 (Function of Degree of Disagreement, FDOD), 通过离散度函数对子序列分布的差异进行度量。当子序列长度一定时, FDOD 方法的计算复杂度随序列规模的增加而呈线性增长, 这一优点使其还适于处理大规模的氨基酸数据^{[10]925}。

本文提出了一种新的 DNA 序列信息表示方法——BB 信息集, 利用不同距离的碱基对的联合概率分布的离散度来表示序列之间的差异。该信息集考虑了 16 种碱基对的联合概率分布, 以及同一种碱基对在不同间隔下的概率分布。实验结果表明该方法是有效的。同时, 分析了信息集变化对结果的影响。当信息集变化时, 相似度较高的序列间的距离值变化很小, 因此, 在运用该方法时不必提取序列的完全信息集, 从而节省了运算时间。

1 相关知识

1.1 完全信息集

设 Σ 为 $\{A, G, C, T\}$ 组成的字符集合, 其中 A, G, C, T 分别表示构成 DNA 序列的 4 种碱基, 则完全测序的 DNA 序列可以表示成 Σ 中的字符组成的字符序列。假设序列 S 的长度为 L , 其长度为 l 的一个子序列是 S 中一条长度为 l 的连续片断。由

收稿日期: 2008-07-22; 修回日期: 2008-09-20。

基金项目: 国家自然科学基金资助项目(60873184); 湖南省自然科学基金资助项目(07JJ5086)。

作者简介: 骆嘉伟(1964-), 女, 湖南长沙人, 教授, 主要研究方向: 生物信息学、数据挖掘; 刘芳(1983-), 女, 湖南娄底人, 硕士研究生, 主要研究方向: 生物信息学; 杨华(1984-), 女, 湖南湘乡人, 硕士研究生, 主要研究方向: 生物信息学。

Σ 中的字母组成的长度为 l 的序列有 4^l 条, 用 p_i^l 表示第 i 条在序列 S 中出现的概率, $p_i^l = n_i^l / (L - l + 1)$, 其中 n_i^l 是这个子序列在序列 S 中出现的次数, 则序列 S 的一个长度为 l 的子序列分布表示为:

$$\mathbf{U}_S^l = (p_1^l, p_2^l, \dots, p_{4^l}^l) \quad (1)$$

对于序列 S , 我们可以得到它的唯一的子序列分布集合:

$$\mathbf{U}_S^l = \{ \mathbf{U}_S^1, \mathbf{U}_S^2, \dots, \mathbf{U}_S^{16} \} \quad (2)$$

这个集合包含了序列 S 的全部结构信息, 我们称之为序列 S 的完全信息集^[9]。

1.2 FDOD 方法

文献[9]提出一种新的信息离散度量方法——FDOD 方法。以子序列分布来表示序列信息时, 离散度函数具有增加性、继承性、存在最大值等很好的数学性质, 该方法已成功用于研究生物进化^[10-11]、蛋白质结构类预测^[12]、SARS 病毒分析^[13]、同源寡聚蛋白质分类^[14-15], 以及 DNA 序列的相似性分析^[16]。

假设 2 条序列 S_1, S_2 在子序列长度为 l 时的子序列分布分别为:

$$\begin{cases} \mathbf{U}_{S_1}^l = (p_{11}^l, p_{21}^l, \dots, p_{4^l}^l) \\ \mathbf{U}_{S_2}^l = (p_{12}^l, p_{22}^l, \dots, p_{4^l}^l) \end{cases} \quad (3)$$

其中: $\sum_{i=1}^{4^l} p_{ik}^l = 1, 1 \leqslant l \leqslant L^k, L^k$ 是序列 S_k 的长度。

这两条序列之间的差异指数可以通过式(4) 来计算:

$$B(\mathbf{U}_{S_1}^l, \mathbf{U}_{S_2}^l) = \sum_{k=1}^2 \sum_{i=1}^{4^l} p_{ik}^l \log \frac{p_{ik}^l}{\sum_{j=1}^{4^l} p_{ij}^l / 2} \quad (4)$$

多条序列 S_1, S_2, \dots, S_t 进行比较时, 其差异指数用式(5) 来计算:

$$B(\mathbf{U}_{S_1}^l, \mathbf{U}_{S_2}^l, \dots, \mathbf{U}_{S_t}^l) = \sum_{k=1}^t \sum_{i=1}^{4^l} p_{ik}^l \log \frac{p_{ik}^l}{\sum_{j=1}^t p_{ij}^l / t} \quad (5)$$

比较序列 S_k 和其他序列 S_1, S_2, \dots, S_t 之间的相似性, 通过式(6) 来计算其差异指数:

$$B_k(\mathbf{U}_{S_k}^l, \mathbf{U}_{S_1}^l, \mathbf{U}_{S_2}^l, \dots, \mathbf{U}_{S_t}^l) = \sum_{i=1}^{4^l} p_{ik}^l \log \frac{p_{ik}^l}{(p_{ik}^l + \sum_{j=1}^t p_{ij}^l) / (t + 1)} \quad (6)$$

我们约定 $0 \times \log \frac{0}{0} = 0, 0 \times \log 0 = 0$ 。差异指数 B 越小, 就认为它们越相似。

但是, 利用 FDOD 方法进行序列相似性分析时, 子序列长度 l 的变化对结果的影响较大, 合适的 l 值通常需要通过训练集来选取^{[15]39-41}, 或者根据经验公式来选取^[17]。通过训练集选取会增加额外的开销, 而经验公式未必适合各种不同的数据集。并且, 随着子序列长度的增加, 子序列分布空间的规模呈指数增长(4^l), 如果选取的子序列长度较长, 计算复杂度将会大大增加。

2 基于信息离散度的 DNA 序列相似性分析方法

2.1 BB 信息集

已测序的 DNA 序列可以表示成 A, G, C, T 组成的字符序列, 则可组成 16 种碱基对, 分别为 $AA, AG, AC, AT, GA, GG, GC, GT, CA, CG, CC, CT, TA, TG, TC, TT$ 。我们用 $n_i^l (1 \leqslant i \leqslant 16)$ 来表示距离为 l 的碱基对在序列中出现的次数。当距离从

1 取到 L 时, 可以得到矩阵 N^l , 见式(7)。令 $total^l = \sum_{i=1}^{16} \sum_{l=1}^L n_i^l$, $p_i^l = n_i^l / total^l$, 得到矩阵 \mathbf{p}^l , 见式(8)。显然有 $\sum_{i=1}^{16} \sum_{l=1}^L p_i^l = 1$ 。

$$N^l = \begin{bmatrix} n_1^1 & n_1^2 & \cdots & n_1^L \\ n_2^1 & n_2^2 & \cdots & n_2^L \\ \vdots & \vdots & & \vdots \\ n_{16}^1 & n_{16}^2 & \cdots & n_{16}^L \end{bmatrix} \quad (7)$$

$$\mathbf{p}^l = \begin{bmatrix} p_1^1 & p_1^2 & \cdots & p_1^L \\ p_2^1 & p_2^2 & \cdots & p_2^L \\ \vdots & \vdots & & \vdots \\ p_{16}^1 & p_{16}^2 & \cdots & p_{16}^L \end{bmatrix} \quad (8)$$

我们称 \mathbf{p}^l 为序列的一个 BB(base-base) 信息集。假设该序列的长度为 L_k , 则当 $L = L_k - 1$ 时, 矩阵 \mathbf{p}^l 包含序列 S 的全部结构信息, 我们称之为 BB 完全信息集。此完全信息集的元素个数为 $(L_k - 1) \times 16$, 而用子序列分布表示的完全信息集的元素个数为 $\sum_{l=1}^{L_k} 4^l$, 显然, 建立 BB 完全信息集比建立子序列完全信息集更易实现, 时间复杂性和空间复杂性都小得多。

2.2 序列间的距离度量

假设 2 条序列 S_1, S_2 的 BB 信息集分别为 $\mathbf{p}_{S_1}^l, \mathbf{p}_{S_2}^l$:

$$\mathbf{p}_{S_1}^l = \begin{bmatrix} p_{11}^1 & p_{11}^2 & \cdots & p_{11}^L \\ p_{21}^1 & p_{21}^2 & \cdots & p_{21}^L \\ \vdots & \vdots & & \vdots \\ p_{161}^1 & p_{161}^2 & \cdots & p_{161}^L \end{bmatrix} \quad (9)$$

$$\mathbf{p}_{S_2}^l = \begin{bmatrix} p_{12}^1 & p_{12}^2 & \cdots & p_{12}^L \\ p_{22}^1 & p_{22}^2 & \cdots & p_{22}^L \\ \vdots & \vdots & & \vdots \\ p_{162}^1 & p_{162}^2 & \cdots & p_{162}^L \end{bmatrix} \quad (10)$$

则两条序列之间的距离用离散度函数(式(11))来计算:

$$B(\mathbf{p}_{S_1}^l, \mathbf{p}_{S_2}^l) = \sum_{k=1}^2 \sum_{i=1}^{16} \sum_{l=1}^L p_{ik}^l \log \frac{p_{ik}^l}{\sum_{j=1}^{16} p_{ij}^l / 2} \quad (11)$$

同样约定 $0 \times \log \frac{0}{0} = 0, 0 \times \log 0 = 0$ 。两条序列的差异

指数 B 越小, 就认为它们越相似。

2.3 相似性分析的方法

利用信息离散度进行 DNA 序列相似性分析的方法如下:

- 1) 输入 M 条 DNA 序列 S_1, S_2, \dots, S_M ;
- 2) 设定参数 L ;
- 3) 根据式(7)、(8), 求出每条序列的 BB 信息集 $\mathbf{p}_{S_k}^l (k = 1, 2, \dots, M)$;
- 4) 利用式(11) 求出 $\mathbf{p}_{S_i}^l$ 和 $\mathbf{p}_{S_j}^l$ 的距离 $D_{ij} (i, j = 1, 2, \dots, M)$, 即序列 S_i 和 S_j 的距离;
- 5) 输出距离矩阵 \mathbf{D} 。

距离矩阵 \mathbf{D} 是一个对角元素为 0 的非负实对称矩阵, 反映了物种间的两两距离。 D_{ij} 越大, 两条序列的相似性越低, 反之则相似性越高。

3 实验与结果

3.1 实验数据

从 NCBI 数据库中下载 DNA 序列数据对所提出的方法

进行实验(<http://www.ncbi.nlm.nih.gov>)。一组是8种H5N1型禽流感病毒的HA片段基因序列,序列长度较短(大约1700bp)(见表1);另一组是11种SARS冠状病毒分离

表1 8种禽流感病毒HA片段

物种	访问号	长度/bp	病毒基因名称
DG12	AY585361	1708	A/duck/Guangdong/12/2000
DG40	AY585374	1708	A/duck/Guangdong/40/2000
DG1681	DQ320886	1698	A/duck/Guangxi/1681/2004
DG1793	DQ320887	1698	A/duck/Guangxi/1793/2004
DH1265	DQ320911	1695	A/duck/Hunan/1265/2005
DH1608	DQ320912	1695	A/duck/Hunan/1608/2005
CY447	DQ095624	1698	A/Chicken/Yunnan/447/05
CH18	DQ320925	1695	A/Chinese pond heron/Hong Kong/18/2005

表2 11种SARS冠状病毒分离株

物种	访问号	长度/bp	来源地
CUHK-Su10	AY282752	29736	香港
HKU-39849	AY278491	29742	香港
CUHK-W1	AY278554	29736	香港
Sin2500	AY283794	29711	新加坡
Sin2677	AY283795	29705	新加坡
Sin2679	AY283796	29711	新加坡
Sin2748	AY283797	29706	新加坡
Sin2774	AY283798	29711	新加坡
TOR2	AY274119	29751	加拿大多伦多
TW1	AY291451	29729	台湾
Urbani	AY278741	29727	美国

表3 取L=30时的距离矩阵

物种	DG12	DG40	DG1681	DG1793	DH1265	DH1608	CY447	CH18
DG12	0.00000	0.02011	0.04381	0.03858	0.05590	0.05121	0.05352	0.04841
DG40	0.02011	0.00000	0.03040	0.02675	0.03279	0.03259	0.03542	0.02667
DG1681	0.04381	0.03040	0.00000	0.00260	0.05260	0.04817	0.05906	0.04697
DG1793	0.03858	0.02675	0.00260	0.00000	0.04789	0.04413	0.05143	0.04288
DH1265	0.05590	0.03279	0.05260	0.04789	0.00000	0.00500	0.02231	0.01167
DH1608	0.05121	0.03259	0.04817	0.04413	0.00500	0.00000	0.02585	0.01273
CY447	0.05352	0.03542	0.05906	0.05143	0.02231	0.02585	0.00000	0.02338
CH18	0.04841	0.02667	0.04697	0.04288	0.01167	0.01273	0.02338	0.00000



图1 邻接法构建禽流感病毒的进化树

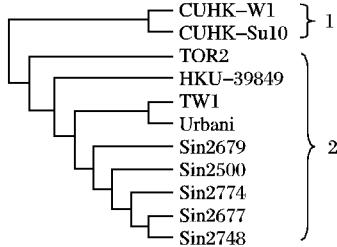


图2 邻接法构建SARS冠状病毒的进化树

当参数L的取值发生变化时,序列之间的距离会如何变化呢?假设信息集为 p^L 时得到的距离矩阵用 $dis(L)$ 表示,定义 $bodong(L)$ 为信息集由 p^{L-1} 变为 p^L 时,得到的两个距离矩

阵的全基因组序列,序列长度较长(大约29700 bp)(见表2)。

3.2 结果与讨论

对表1中的数据,取参数L的值为30,得到8种病毒的距离矩阵,如表3(因为得到的距离值较小,表3列出的数值为原距离值的100倍)。由表3可知,相似度较高的物种对为(DG12,DG40),(DG1681,DG1793),(DH1265,DH1608),(DH1265,CH18),(DH1608,CH18)。

利用表3给出的距离矩阵,用PHYLIP软件包中的neighbor.exe构造物种进化树(图1)。由图1可知,同一年份相同地域的同种动物身上采集的病毒聚在一起。如果利用CLUSTALX软件进行多序列比对后,采用Kimura两参数模型得到距离矩阵,再用neighbor.exe构造进化树,我们发现这两种方法得到的进化树结构完全相同。文献[18]利用基于DNA序列3D图形表示的方法也得到了相同的结果。

对表2中的数据,取L=100,得到距离矩阵后,用PHYLIP软件包中的neighbor.exe构造物种进化树,如图2。由文献[13,19]可知,第2组病毒是与造成香港Metrople饭店疫情的SARS冠状病毒相联系的,第1组则与其没有已知的联系。TOR2和HKU-39849与Metrople饭店疫情的SARS冠状病毒是间接联系^[19,174],因而与第2组中其他病毒分离株相距较远。本文方法得到的距离矩阵对应的进化树与这些已知信息非常相符。

阵中对应元素相减的绝对值之和,即:

$$bodong(L) = \sum |dis(L) - dis(L-1)| \quad (12)$$

图3为DG12与其他7种病毒的距离随L的变化曲线。图4为**bodong(L)**的值随L的变化曲线。由图3可知,物种之间的距离值并不随着L的增大而单调递增或递减。当L取值较小时,由于BB信息集包含的信息较少,距离值的变化较大。随着L值的增大,距离值的变化趋于平缓。图4也证明了这一点。特别是,对于相似度极高的序列,如DG12和DG40,当参数L变化时,其距离值几乎构成一条水平直线(见图3)。所以,尽管L值越大, p^L 包含的信息越多,但我们并不需要对整个BB完全信息集进行分析,根据序列间的距离值变化选取合适的L值即可。

4 结语

本文提出了一种利用不同距离的碱基对的联合概率分布建立信息集的方法,并利用信息离散度函数对DNA序列进行相似性分析。这种方法最突出的特点是代表序列特征的信息

集的建立不带有主观因素,因而能比较客观地反映生物序列间的关系。与 FDOD 方法相比,它简化了信息集的建立,并且对于相似度较高的序列,信息集的变化对结果的影响很小,具有实现简单、空间复杂度和时间复杂度小的优点。本文方法可以对 DNA 序列的相似性进行有效分析,其计算复杂度随序列长度的增加呈线性增长,因此可用于处理大规模数据。这种方法可以用于序列比较、物种分类、生物进化关系研究等领域,以及多序列的比较问题。

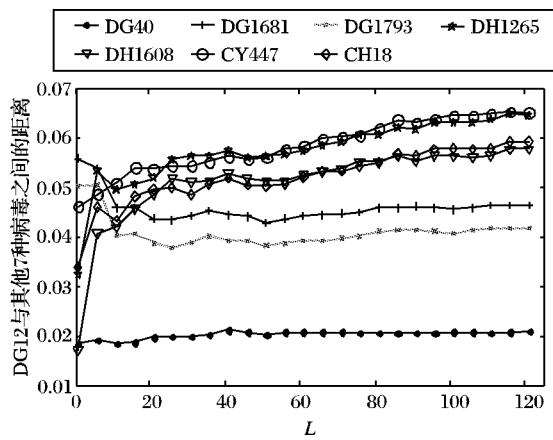


图 3 DG12 与其他 7 种病毒的距离随 L 的变化曲线

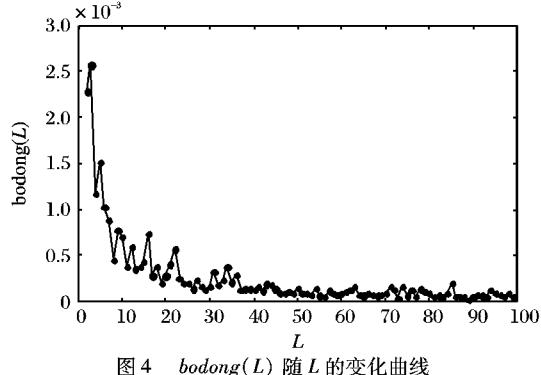


图 4 $bodong(L)$ 随 L 的变化曲线

参考文献:

- [1] 郝柏林, 张淑善. 生物信息学手册 [M]. 上海: 上海科学技术出版社, 2000: 184 - 199.
- [2] 张永, 王瑞. 生物信息学中的序列比对算法 [J]. 电脑知识与技术, 2008(1): 181 - 184.
- [3] ZHANG R, ZHANG C T. Z curves: An intuitive tool for visualizing and analyzing the DNA sequences [J]. Journal of Biomolecular Structure and Dynamics, 1994, 11(4): 767 - 782.
- [4] RANDIC M, VRACKO M, NANDY A, et al. On 3-D graphical representation of DNA primary sequence and their numerical characterization [J]. Chemical Information and Computer Sciences, 2000, 40(1): 1235 - 1244.
- [5] LIAO B, WANG T M. Analysis of similarity/ dissimilarity of DNA sequences based on 3-D graphical representation [J]. Journal of Molecular Structure: THEOCHEM, 2004, 388: 195 - 200.
- [6] ZHAO H Q, TONG R F. PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization [J]. Chemical Physics Letters, 2007, 442: 434 - 440.
- [7] FARACH M, NOORDEWIJER M, SAVARI S, et al. On the entropy of DNA: Algorithm and measurements based on memory and rapid convergence [C]// Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms. San Francisco, 1995: 48 - 57.
- [8] HARID A, WEBER B, OLMSTED J. On the validity of Shannon-information calculations for Molecular biological sequences [J]. Journal of Theoretical Biology, 1990, 147: 235 - 254.
- [9] FANG W W, ROBERTS F S, MA Z R. A measure of discrepancy of multiple sequences [J]. Information Science, 2001, 137: 75 - 102.
- [10] 张文, 唐焕文, 方伟武, 等. 基于全蛋白质组的微生物系统发育树构建 [J]. 大连理工大学学报, 2005, 45(6): 925 - 930.
- [11] WANG J H, FANG W W, LING L X, et al. Gene's functional arrangement as a measure of the phylogenetic relationships of microorganisms [J]. Journal of Biological Physics, 2002, 28: 55 - 62.
- [12] JIN L X, FANG W W, TANG H W. Prediction of protein structural classes by a new measure of information discrepancy [J]. Computational Biology and Chemistry, 2003, 27: 373 - 380.
- [13] 齐震, 狐昱, 李蔚等. 基于全基因组比较的 SARS 冠状病毒种系进化分析 [J]. 科学通报, 2003, 48(12): 1242 - 1245.
- [14] SONG J, TANG H W. Accurate classification of homodimeric vs other homooligomeric proteins using a new measure of information discrepancy [J]. Journal of Chemical Information and Computer Sciences, 2004, 44: 1324 - 1327.
- [15] 宋杰, 唐焕文. 基于一种新的信息离散性度量方法的同源寡聚蛋白质分类 [J]. 数学的实践与认识, 2007, 37(8): 36 - 42.
- [16] 宋杰. 基于离散度函数的 DNA 序列的相似性分析 [J]. 计算机与应用化学, 2007, 6(24): 729 - 733.
- [17] LI W, FANG W W, LING L J, et al. Phylogeny based on whole genome as inferred from complete information set analysis [J]. Journal of Biological Physics, 2002, 28(3): 439 - 447.
- [18] LUO J W, ZHANG X Z. New method for constructing phylogenetic tree based on 3D graphical representation [C]// The 1st International Conference on Bioinformatics and Biomedical Engineering. Wuhan: IEEE Press, 2007: 324 - 327.
- [19] RUAN Y J, WEI C L, EE L A, et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection [J]. The Lancet, 2003, 361(9371): 1779 - 1785.

(上接第 151 页)

通过本模型中所提出的信任建立和更新策略,使得实体在交易之前充分考虑普适环境下实体的信任证据,同时对即将交易的实体进行全面的信任评估,从而能够做出正确的交易决策,提高了系统的安全性。下一步的研究工作方向是在本模型基础上给出更加优化的函数来保持模型中信任值计算的准确度。此外还需要在模型中增加相应的机制来识别协同作弊恶意节点并对其进行惩罚,以保障整个普适系统的安全性和稳定性。

参考文献:

- [1] BLAZE M, FEIGENBAUM J, LACY J. Decentralized trust management [C]// Proceedings of the 17th symposium on Security and Privacy. Oakland: IEEE Computer Society Press, 1996: 164 - 173.

- [2] JOSANG A. Trust management for E-Commerce [C]// Proceedings of the Virtual Banking. Heidelberg: Springer Berlin Press, 2000.
- [3] 郭亚军, 王亮, 洪帆, 等. 基于信任的普适计算的动态授权模型 [J]. 华中科技大学学报: 自然科学版, 2007(8): 70 - 73.
- [4] SHAND B, DIMMOEK N, BACON J. Trust for ubiquitous, transparent collaboration [C]// Proceedings of the 1st Annual IEEE Conference on Pervasive Computing and Communications. Texas: Springer Netherlands Press, 2003: 153 - 160.
- [5] XIU D X, LIU Z Y. A dynamic trust model for pervasive computing environments [C]// Proceedings of the 4th Annual Security Conference. Las Vegas: IEEE Computer Society Press, 2005.
- [6] 常俊胜, 王怀民, 尹刚. DyTrust: 一种 P2P 系统中基于时间帧的动态信任模型 [J]. 计算机学报, 2006, 29(8): 1301 - 1307.